Discriminant Incoherent Component Analysis

Christos Georgakis, *Student Member, IEEE,* Yannis Panagakis, *Member, IEEE,* and Maja Pantic, *Fellow, IEEE*

***Abstract*—Face images convey rich information which can be perceived as a superposition of low-complexity components associated with attributes, such as facial identity, expressions and activation of facial action units (AUs). For instance, low-rank components characterizing neutral facial images are associated with identity, while sparse components capturing non-rigid deformations occurring in certain face regions reveal expressions and action unit activations. In this paper, the Discriminant Incoherent Component Analysis (DICA) is proposed in order to extract low-complexity components corresponding to facial attributes, which are mutually incoherent among different classes (e.g., identity, expression, AU activation) from training data, even in the presence of gross sparse errors. To this end, a suitable optimization problem, involving the minimization of nuclear- and ℓ_1-norm, is solved. Having found an ensemble of class-specific incoherent components by the DICA, an unseen (test) image is expressed as a group-sparse linear combination of these components, where the non-zero coefficients reveal the class(es) of the respective facial attribute(s) that it belongs to. The performance of the DICA is experimentally assessed on both synthetic and real-world data. Emphasis is placed on face analysis tasks, namely joint face and expression recognition, face recognition under varying percentages of training data corruption, subject-independent expression recognition, and action unit detection by conducting experiments on 4 datasets. The proposed method outperforms all the methods that is compared to in all tasks and experimental settings.**

***Index Terms*—Discriminant Incoherent Component Analysis, Incoherent Subspaces, Sparse-based Representation Classification, Low-rank, Sparsity**

I. Introduction

FACE analysis has been an active research topic over the last thirty years. Human face is a rich source of information consisting of several components which are related to attributes associated with facial identity, emotional expression and activation of action units (AUs). These components are characterized by specific structures which can assist the semantic interpretation of content in the visual stream. For instance, facial expressions manifest themselves through *sparse* non-rigid deformations occurring in certain face regions [1], [2], while images depicting the neutral face of the same person are expected to be highly correlated and thus drawn from a *low-rank* subspace. Consequently, the extraction of such features of low-complexity (i.e., exhibiting low-rank or sparse structure) is essential for accurate face and expression recognition.

Machine learning approaches to face recognition primarily aim to extract discriminant features that are invariant to pose, expression and illumination variations in order to train classifiers. To achieve this, subspace analysis methods such as Eigenfaces [3], Fisherfaces [4], Laplacianfaces [5], Locally Linear Embedding [6], [7] and Isomap [8] aim at feature extraction, based on the assumption that the high-dimensional observed faces live in a low-dimensional space. However, those methods are susceptible to non-Gaussian, gross contamination in the data (e.g., occlusions). A partial remedy to this issue has been provided by Robust Principal Component Analysis [9] and other similar approaches (e.g., [10]–[13]), whose building block is the decomposition of face imagery into a low-rank part and an error term accounting for sparse corruptions, occlusions, and outliers. Subspace learning on Image Gradient Orientations (IGO) [14] also alleviates the problem of illumination- and corruption-related noise without significantly increasing the computational complexity. On the other hand, Sparse Representation-based Classification (SRC) [15] has boosted the development of methods that focus more on face representation. The main assumption of SRC is that an unseen (test) image can be represented as a sparse linear combination of the training face images or discriminative, noise-free atoms [16]–[22].

Most works on facial expression recognition focus on "message judgement" – classifying observed facial expressions in terms of emotions or other messages (e.g. pain, interest, stance, accent) [23]–[26]. Various features have been employed for this task, including Gabor features (e.g. [27]) and Local Binary Patterns (LBP) [28], [29]. SRC has been shown to be efficient also for recognition of emotional expression [27], [30]–[33]. Other works in the field focus on "sign judgement" – classifying observed facial expressions in terms of facial muscle activations (AUs) that produced the observed expression [1], [2], [34]. These atomic facial actions correspond to all visually discernible facial movements and can be measured according to the facial action coding system (FACS) [35].

Face and facial expression recognition, despite being two intertwined tasks within the context of face analysis, have hitherto been targeted jointly by just a few works. Vasilescu and Terzopoulos [36] employ an extension of Singular Value Decomposition (SVD) to tensors to uncover subspaces generating different faces, expressions, viewpoints, and illuminations. Another SVD-based work is [37], where the proposed Higher-Order SVD is used to learn the mapping between persons and expressions, which is subsequently utilized to perform facial expression decomposition. Recently, Taheri et al. [11] combine RPCA [9] and K-SVD [38] to construct one identity and one expression dictionary, which are in turn fed into a SRC-like framework for joint face and expression recognition.

The fundamental constraint of the above mentioned methods for face analysis is that the training data is often assumed to be noise-free. That is, they are collected under

C. Georgakis is with the Department of Computing, Imperial College London, London, UK, (e-mail: christos.georgakis@imperial.ac.uk).

Y. Panagakis is with the Department of Computing, Imperial College London, London, UK, (e-mail: i.panagakis@imperial.ac.uk).

M. Pantic is with the Department of Computing, Imperial College London, London, U.K., and also with EEMCS, University of Twente, Enschede, The Netherlands (e-mail: m.pantic@imperial.ac.uk).

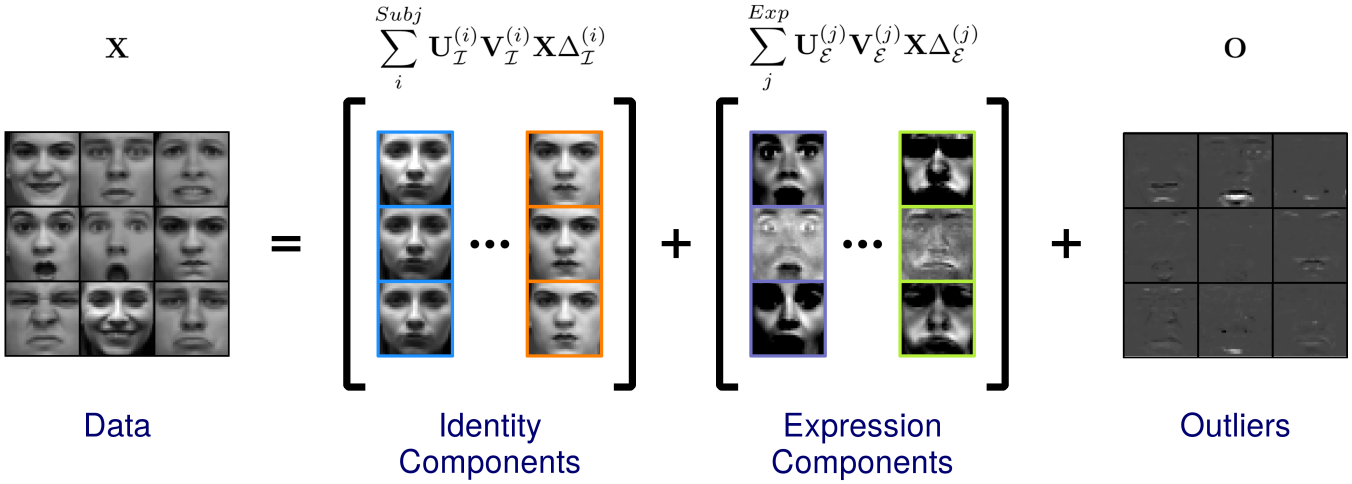


Fig. 1. The proposed Discriminant Incoherent Component Analysis (DICA), as applied to the multi-label setting of joint face and expression recognition. The data matrix $\mathbf{X}$ containing expressive face images is expressed as a superposition of identity- and expression-specific mutually incoherent components, under the assumption of possible gross errors (outliers).

well controlled conditions in terms of illumination and pose variations and they do not contain occlusions or disguise. Consequently, the aforementioned methods are not applicable in practical scenarios when both training and test data are contaminated by gross non-Gaussian noise and corruptions (e.g., occlusions and disguise). Moreover, the majority of these works approach the tasks of face and expression recognition separately rather than within a joint framework.

To alleviate the aforementioned drawbacks and motivated by recent advances in robust subspace learning [13], [39]–[43], we propose the Discriminant Incoherent Component Analysis (DICA) in order to decompose training facial images into a superposition of class-specific structured and mutually incoherent components accounting for identity, emotional expression or AUs in the presence of gross but sparse non-Gaussian corruptions. In other words, we model expressive faces as expressionless faces capturing the identity, superimposed by sparse images of non-rigid deformations corresponding to facial expressions, plus sparse components corresponding sparse errors of large magnitude, which cannot be explained by labels. To learn such a decomposition, we impose low-rank constraints on the components capturing the face's identity and sparsity constraints to those related to expressions. The proposed model can be also used to recover more localized sparse components related to AUs. Having found an ensemble of class-specific incoherent components, a test image is expressed as a group-sparse linear combination of these components with non-zero coefficients corresponding to the identity and expression class that the test sample belongs to. Overall, this discriminative representation furnished by the DICA proves efficient for the related classification tasks.

The contributions of this paper are as follows:

1) The DICA provides a generic method to decompose data into class-specific structured and incoherent components, and a sparse matrix accounting for outliers.
2) An efficient Alternating-Directions Method of Multipliers (ADMM)-based algorithm is presented that can solve suitable optimization problems for the DICA, according to the desirable component structure.
3) A dictionary-based classification framework is proposed, according to which a test sample is collaboratively represented via class-specific components extracted by the DICA.

The performance of the DICA is assessed by conducting experiments on joint face and expression recognition, face recognition under varying percentages of training data corruption, subject-independent expression recognition under varying illumination conditions during training, and facial action unit detection, using 4 datasets. The proposed method outperforms the methods that is compared to in all the aforementioned tasks.

The remainder of the paper is as follows. In Section II, the DICA and its algorithmic framework are detailed. A dictionary-based framework for classification via the DICA is described in section III. The performance is assessed experimentally on both synthetic and real-world data in Section IV. Section V concludes the paper and gives insight for future research directions.

Notations: Matrices (vectors) are denoted by uppercase (lowercase) boldface letters , e.g., $\mathbf{A}, \mathbf{B}, (\mathbf{a}, \mathbf{b})$. $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The ith element of vector $\mathbf{x}$ is denoted as x_i, while the ith column of matrix $\mathbf{X}$ is denoted as $\mathbf{x_i}$. For the set of real numbers, the symbol $\mathbb{R}$ is used. We refer to a set of N real matrices of varying dimensions as $\{\mathbf{X}^{(n)} \in \mathbb{R}^{p_n \times q_n}\}_{n=1}^{N}$. Regarding vector norms, $\|\mathbf{x}\| = \sqrt{\sum_i x_i^2}$ denotes the Euclidean norm. Regarding matrix norms, $\|\mathbf{X}\|_*$ denotes the nuclear norm, which equals the sum of singular values, while $\|\mathbf{X}\|$ denotes the spectral norm, which equals the largest singular value. $\|\mathbf{X}\|_1 = \sum_i \sum_j |x_{ij}|$ is the element-wise matrix ℓ_1-norm, and $\|\mathbf{X}\|_F = \sqrt{\sum_i \sum_j x_{ij}^2} = \sqrt{\text{tr}(\mathbf{X}^T\mathbf{X})}$ is the Frobenius norm, with $\text{tr}(\cdot)$ denoting the trace of a square matrix. Finally, $\lambda_{\max}[\mathbf{X}]$ denotes the largest eigenvalue of a square matrix $\mathbf{X}$.

II. Discriminant Incoherent Component Analysis

In this section, the DICA is described along with its solver.

A. Problem Statement

The goal of the DICA is to robustly learn components from training samples that 1) are discriminant and exhibit low-complexity structures (e.g., low-rank or sparsity) associated with facial attributes, 2) are mutually incoherent among different classes, and 3) facilitate the classification of test samples by means of sparse representation.

Let $\mathbf{x} \in \mathbb{R}^d$ be a vectorized expressive face image and $\boldsymbol{l} \in \{0,1\}^{n_c}$ the label vector associated with it, whose non-zero elements are those corresponding to the identity and expression class it belongs to (n_c denotes the total number of classes). We seek to decompose $\mathbf{x}$ as a sum of n_c class-specific components $\mathbf{y}^{(i)} \in \mathbb{R}^d$, capturing the discriminant characteristics of each class. Thus, $\mathbf{x}$ is expressed as

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{y}^{(i)} \tag{1}$$

We assume that each class-specific component $\mathbf{y}^{(i)}$ lies in a linear orthonormal subspace spanned by $\mathbf{U}^{(i)} \in \mathbb{R}^{d \times m^{(i)}}$, and $\mathbf{V}^{(i)} \in \mathbb{R}^{m^{(i)} \times d}$ denotes the projection matrix that embeds $\mathbf{x}$ onto the $m^{(i)}$-dimensional space, while also preserving the structure (e.g., low-rank or sparsity) related to the class-specific attribute. Therefore, $\mathbf{y}^{(i)}$ is written as

$$\mathbf{y}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{x}\,, \tag{2}$$

Following [44] and [13], the above mentioned formulation enables us to impose a specific structure on the projection spaces $\mathbf{V}^{(i)}$, by minimizing a suitable structure-inducing norm $\|\mathbf{V}^{(i)}\|_{(\cdot)}$; this is either the nuclear norm [45] which imposes low-rank on the projection spaces corresponding to facial identities, or the ℓ_1-norm [46] which enables to learn sparse projections for facial expressions or AUs. By incorporating (2) into (1), $\mathbf{x}$ is written as

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{x}\,, \tag{3}$$

Clearly, to perfectly disentangle the class-specific components $\mathbf{y}^{(i)}$ (i.e., to ensure the identifiability of (1)), the column spaces that they are stemming from should be mutually incoherent, that is $\mathbf{U}^{(i)^T}\mathbf{U}^{(j)} = \mathbf{0}$ for $i \neq j$. We observe that Equation (3), combined with the mutual incoherence property $\mathbf{U}^{(i)^T}\mathbf{U}^{(j)} = \mathbf{0}$ for $i \neq j$, entails $\mathbf{U}^{(i)^T} \simeq \mathbf{V}^{(i)}$ for $i = 1, 2, \ldots, n_c$. In other words, matrices $\mathbf{U}^{(i)^T}$ and $\mathbf{V}^{(i)}$ are proportional for every class i. This further entails that $\mathbf{U}^{(i)^T}\mathbf{U}^{(j)} = \mathbf{0}$ is equivalent to $\mathbf{V}^{(i)}\mathbf{V}^{(j)^T} = \mathbf{0}$ for $i \neq j$.

To account also for the possible presence of facial aspects that cannot be explained by labels, including outliers and gross corruptions, we include the additive term $\mathbf{o} \in \mathbb{R}^d$ in the decomposition (3), which is written as

$$\mathbf{x} = \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{x} + \mathbf{o}\,, \tag{4}$$

Having found the decomposition (4), the representation vector $[(\mathbf{V}^{(1)}\mathbf{x})^T, (\mathbf{V}^{(2)}\mathbf{x})^T, \cdots, (\mathbf{V}^{(n_c)}\mathbf{x})^T]^T$ is expected to be group-sparse, with non-zero elements corresponding to the class(es) the sample $\mathbf{x}$ belongs to.

The DICA learns the reconstruction matrices $\{\mathbf{U}^{(i)}\}_{i=1}^{n_c}$ and projection matrices $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$ by employing the training matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$ which contains in its columns the vectorized training face images, with d being the dimensionality of each image and N the number of training observations. Let us denote by $\mathbf{X}_{\mathcal{S}^{(i)}} \in \mathbb{R}^{d \times N}$ the column-sparse matrix whose non-zero columns are the columns of $\mathbf{X}$ with label i. Therefore, with the set $\mathcal{W} = \{\{\mathbf{U}^{(i)} \in \mathbb{R}^{d \times m^{(i)}}\}_{i=1}^{n_c}, \{\mathbf{V}^{(i)} \in \mathbb{R}^{m^{(i)} \times d}\}_{i=1}^{n_c}, \mathbf{O} \in \mathbb{R}^{d \times N}\}$ containing all the unknown variables, the DICA solves

$$\begin{aligned} \arg\min_{\mathcal{W}} \; & \lambda^{(i)} \sum_{i=1}^{n_c} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i \neq j} \|\mathbf{V}^{(i)}\mathbf{V}^{(j)^T}\|_F^2 + \lambda_1 \|\mathbf{O}\|_1\,, \\ \text{s.t.} \quad & i) \quad \mathbf{X} = \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} + \mathbf{O}\,, \\ & ii) \quad \mathbf{U}^{(i)^T}\mathbf{U}^{(i)} = \mathbf{I}\,, \quad i = 1, 2, \ldots, n_c\,, \end{aligned} \tag{5}$$

where the structure-inducing norm $\|\mathbf{V}^{(i)}\|_{(\cdot)}$ is either the nuclear norm for face-specific projections or the ℓ_1-norm for expression-specific and AU-specific projections. The term $\sum_{i \neq j} \|\mathbf{V}^{(i)}\mathbf{V}^{(j)^T}\|_F^2$ induces mutual incoherence among the projection spaces and $\mathbf{O} \in \mathbb{R}^{d \times N}$ denotes the outlier matrix accounting for components that cannot be explained by the summand containing the class-specific reconstructions. The positive parameters $\lambda^{(i)}$, η, and λ_1 control the norm imposed on $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$, the mutual incoherence for all component pairs, and the sparsity of outliers $\mathbf{O}$, respectively.

In Fig. 1, one can see how the proposed DICA is applied to the multi-label scenario of joint face and expression recognition. In that case, each training image is characterized by two labels, one for identity and the other for expression. The data matrix $\mathbf{X}$, containing the vectorized training images, is accordingly represented as a superposition of discriminant and mutually incoherent class-specific components (low-rank for identity and sparse for expression), plus an outlier matrix $\mathbf{O}$ accounting for unbounded sparse errors.

B. Alternating-Direction Method-Based Algorithm

The Alternating-Directions Method of Multipliers (ADMM) [47] is employed hereby to solve (5). The (partial) augmented Lagrangian function for (5) is defined as:

$$\begin{aligned} \mathcal{L}(\mathcal{W}, \mathbf{Y}, \mu) = {} & \lambda^{(i)} \sum_{i=1}^{n_c} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i \neq j} \|\mathbf{V}^{(i)}\mathbf{V}^{(j)^T}\|_F^2 \\ & + \lambda_1 \|\mathbf{O}\|_1 + \mathrm{tr}\left(\mathbf{Y}^T\left(\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\right)\right) \\ & + \frac{\mu}{2}\|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\|_F^2\,, \end{aligned} \tag{6}$$

Algorithm 1 ADMM solver for the DICA (5)

Input: Data: $\mathbf{X} \in \mathbb{R}^{d\times N}$. Parameters: $\lambda^{(i)}$, η, λ_1, and $\{m^{(i)}\}_{i=1}^{n_c}$.

Normalize each column of $\mathbf{X}$ to unit ℓ_2-norm.
Initialize: Set $\{\{\mathbf{U}^{(i)}[0]\}, \{\mathbf{V}^{(i)}[0]\}\}_{i=1}^{n_c}$, $\mathbf{O}[0]$, $\mathbf{Y}[0]$ to zero matrices. Set $\mu[0] = 1/\|\mathbf{X}\|$, $\rho = 1.1$, $\mu_{\max} = 10^{10}$.
while not converged **do**
  for $i = 1 : n_c$ **do**
    Calculate $L = 1.02\lambda_{\max}\left[\mu[t]\mathbf{X}_{\mathcal{S}^{(i)}}\mathbf{X}_{\mathcal{S}^{(i)}}^T + 2\eta\sum_{j\neq i}\mathbf{V}^{(j)}[t]^T\mathbf{V}^{(j)}[t]\right]$.
    if $\mathbf{V}^{(i)}$ is associated with nuclear norm **then**
      $\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{D}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - L^{-1}\nabla f(\mathbf{V}^{(i)}[t])\right]$.[1]
    else if $\mathbf{V}^{(i)}$ is associated with ℓ_1-norm **then**
      $\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{S}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - L^{-1}\nabla f(\mathbf{V}^{(i)}[t])\right]$.
    end if
    $\mathbf{U}^{(i)}[t+1] \leftarrow \mathcal{P}\left[\left(\mathbf{X} - \sum_{j\neq i}\mathbf{U}^{(j)}[t]\mathbf{V}^{(j)}[t+1]\mathbf{X}_{\mathcal{S}^{(j)}} - \mathbf{O}[t] + \mu[t]^{-1}\mathbf{Y}[t]\right)\left(\mathbf{V}^{(i)}[t+1]\mathbf{X}_{\mathcal{S}^{(i)}}^T\right)\right]$.
  end for
  $\mathbf{O}[t+1] \leftarrow \mathcal{S}_{\lambda_1/\mu[t]}\left[\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}[t+1]\mathbf{V}^{(i)}[t+1]\mathbf{X}_{\mathcal{S}^{(i)}} + \mu[t]^{-1}\mathbf{Y}[t]\right]$.
  Update the Lagrange multiplier by $\mathbf{Y}[t+1] \leftarrow \mathbf{Y}[t] + \mu[t]\left(\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}[t+1]\mathbf{V}^{(i)}[t+1]\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}[t+1]\right)$.
  Update μ by $\mu[t+1] = \min(\rho\cdot\mu[t], \mu_{\max})$.
end while

Output: $\{\mathbf{U}^{(i)} \in \mathbb{R}^{d\times m^{(i)}}, \mathbf{V}^{(i)} \in \mathbb{R}^{m^{(i)}\times d}\}_{i=1}^{n_c}$, $\mathbf{O} \in \mathbb{R}^{d\times N}$.

where μ is a positive parameter and $\mathbf{Y} \in \mathbb{R}^{d\times N}$ is the Lagrange multiplier related to the linear constraint in (5).

At each iteration, (6) is minimized with respect to each variable in $\mathcal{W}$ in an alternating fashion and, subsequently, the Lagrange multiplier $\mathbf{Y}$ and parameter μ are updated. The iteration index is denoted herein by t. The notation $\mathcal{L}(\mathbf{U}^{(i)}, \mathbf{Y}[t], \mu[t])$ is used to denote the solution stage in which all other variables but $\mathbf{U}^{(i)}$ are kept fixed, and similarly for the other unknown variables. Thus, given the variables $\mathcal{W}[t]$, the Lagrange multiplier $\mathbf{Y}[t]$ and the parameter $\mu[t]$ at iteration t, the updates of ADMM are calculated as follows.

Update the primal variables:

$$\begin{aligned}\mathbf{U}^{(i)}[t+1] &= \arg\min_{\mathbf{U}^{(i)}} \mathcal{L}(\mathbf{U}^{(i)}, \mathbf{Y}[t], \mu[t]) \\ &\quad \text{s.t.} \quad \mathbf{U}^{(i)^T}\mathbf{U}^{(i)} = \mathbf{I}, \quad i = 1, 2, \ldots, n_c \\ &= \arg\min_{\mathbf{U}^{(i)}} \frac{\mu[t]}{2}\|\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1}\mathbf{Y}\|_F^2 \\ &\quad \text{s.t.} \quad \mathbf{U}^{(i)^T}\mathbf{U}^{(i)} = \mathbf{I}, \quad i = 1, 2, \ldots, n_c\end{aligned} \tag{7}$$

$$\begin{aligned}\mathbf{V}^{(i)}[t+1] &= \arg\min_{\mathbf{V}^{(i)}} \mathcal{L}(\mathbf{V}^{(i)}, \mathbf{Y}[t], \mu[t]) \\ &= \arg\min_{\mathbf{V}^{(i)}} \lambda^{(i)}\|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta\sum_{i\neq j}\|\mathbf{V}^{(i)}\mathbf{V}^{(j)^T}\|_F^2 \\ &+ \frac{\mu[t]}{2}\|\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1}\mathbf{Y}\|_F^2 \\ &= \arg\min_{\mathbf{V}^{(i)}} \lambda^{(i)}\|\mathbf{V}^{(i)}\|_{(\cdot)} + f(\mathbf{V}^{(i)}), \quad i = 1, 2, \ldots, n_c\end{aligned} \tag{8}$$

Algorithm 2 Framework for face/expression recognition.

Input: Data: training set $\mathbf{X} \in \mathbb{R}^{d\times N}$, query image $\mathbf{y} \in \mathbb{R}^{N\times 1}$. Parameters: λ_{Lasso}.

Normalize each column of $\mathbf{X}$ to unit ℓ_2-norm.
Compute low-rank matrices $\{\mathbf{A}^{(i)}\}_{i=1}^{n_c}$ by performing RPCA [9] on each class-specific sub-matrix $\mathbf{X}^{(i)}$.
Initialize: For each subspace $i \in \{1, 2, \ldots, n_c\}$, set $\mathbf{U}^{(i)}[0] = \mathbf{M}^{(i)}$, and $\mathbf{V}^{(i)}[0] = \mathbf{M}^{(i)^T}$, where $\mathbf{A}^{(i)} = \mathbf{M}^{(i)}\mathbf{\Sigma}\mathbf{N}^{(i)^T}$ is the skinny SVD of $\mathbf{A}^{(i)}$.
Calculate $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$ according to Algorithm 1, using the nuclear- (ℓ_1-) norm in Problem (5) for face (expression) recognition.
Form dictionary $\mathbf{D} = \left[\mathbf{D}^{(1)}, \mathbf{D}^{(2)}, \ldots, \mathbf{D}^{(n_c)}\right]$, with $\mathbf{D}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}$, $i \in \{1, 2, \ldots, n_c\}$.
Normalize each column of $\mathbf{D}$ to unit ℓ_2-norm.
Perform SRC: $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|^2 + \lambda_{Lasso}\|\boldsymbol{\alpha}\|_1$.
for $i = 1 : n_c$ **do**
  $err(i) = \|\mathbf{y} - \mathbf{D}\delta^{(i)}(\hat{\boldsymbol{\alpha}})\|$.
end for
$i^* \leftarrow \arg\min_{i\in\{1,2,\ldots,n_c\}} err(i)$.

Output: subject (expression) label i^*.

$$\begin{aligned}\mathbf{O}[t+1] &= \arg\min_{\mathbf{O}} \mathcal{L}(\mathbf{O}, \mathbf{Y}[t], \mu[t]) \\ &= \arg\min_{\mathbf{O}} \lambda_1\|\mathbf{O}\|_1 \\ &+ \frac{\mu[t]}{2}\|\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1}\mathbf{Y}\|_F^2\end{aligned} \tag{9}$$

Update the Lagrange Multiplier:

$$\mathbf{Y}[t+1] = \mathbf{Y}[t] + \mu[t]\left(\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\right) \quad (10)$$

Equations (7)-(9) are solved by means of the operators and Lemmas that are introduced next. We begin by defining the shrinkage operator [9] as $\mathcal{S}_\tau[a] = \text{sgn}(a)\max(|a| - \tau, 0)$, whose matrix version is obtained by applying it element-wise. Also, if $\mathbf{A} = \mathbf{M}\mathbf{\Sigma}\mathbf{N}^T$ denotes the SVD of a matrix $\mathbf{A}$, the singular value thresholding operator (SVT) is defined as in [48]: $\mathcal{D}_\tau[\mathbf{A}] = \mathbf{M}\mathcal{S}_\tau[\mathbf{\Sigma}]\mathbf{N}^T$. Based again on the SVD of $\mathbf{A}$, the Procrustes operator is defined as $\mathcal{P}[\mathbf{A}] = \mathbf{M}\mathbf{N}^T$ and solves the problem in the following Lemma.

Lemma 1 [44]: *The constrained minimization problem:*

$$\arg\min_{\mathbf{B}} \|\mathbf{A} - \mathbf{B}\|_F^2 \quad \text{s.t.} \quad \mathbf{B}^T\mathbf{B} = \mathbf{I} \quad (11)$$

has a closed-form solution given by $\mathbf{P} = \mathcal{P}[\mathbf{A}]$.

The solution of (8) is presented in detail in the Appendix and is based on the SVT (shrinkage) operator when the nuclear- (ℓ_1-) norm is employed for the component $\mathbf{V}^{(i)}$. Moreover, the minimizer of (9) is based on the shrinkage operator. Finally, (7) is solved as in Lemma 1. The ADMM-based solver of (5) is wrapped up in Algorithm 1. For all experiments presented herein, Algorithm 1 is terminated when $\|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}\|_F / \|\mathbf{X}\|_F < 10^{-7}$, or when 1000 iterations are reached.

Computational Complexity and Convergence: In the case where the nuclear norm is enforced on $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$, the cost of each iteration in Algorithm 1 is mainly associated with the calculation of the SVT operator in Step 7. Hence, each iteration has a complexity equal to that of SVD, i.e., $\mathcal{O}(\max(d^2N, dN^2))$. In the case where the ℓ_1-norm is used, the shrinkage operator becomes the most time-consuming calculation, thus entailing linear complexity $\mathcal{O}(dN)$. As far as convergence of Algorithm 1 is concerned, the convergence of the ADMM to local minima has not been proved for the cases where the latter is adopted to solve non-convex problems [47], [49]. A systematic convergence proof does not fall within the scope of this paper, yet for proof of the weak convergence of Algorithm 1 one can follow the approach in [50]. Nonetheless, the experiments in Section IV serve as a testament to the guaranteed convergence of Algorithm 1.

III. DICA-based Classification

In this section, a dictionary-based framework built upon the DICA (5) is proposed. This can be tailored accordingly to cope with either a single- or a multi-label scenario. Herein, the framework is presented for the problems of face and expression recognition, viewed either as separate single-label tasks or jointly within a multi-label setting. For the multi-label scenario, an extension of our framework, which can deal with the facial action unit detection task, is also described.

[1] f is the smooth differentiable part of the minimizer (8).

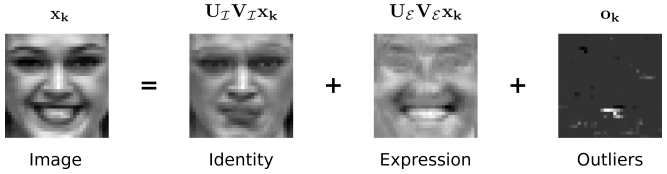


Fig. 2. Decomposition of an expressive image from the CK+ Dataset into an identity component, an expression component and a sparse error term accounting for outliers, as produced by the DICA.

A. Single-Label Case: Face/Expression Recognition

Suppose each column $\mathbf{x_n}$ of our training data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$ represents a vectorized image, with subject (expression) label $i \in \{1, 2, \ldots, n_c\}$, where n_c equals the number of subjects (expressions). Let us also denote by $\mathbf{X}^{(i)} \in \mathbb{R}^{d\times n^{(i)}}$ the matrix that is composed of the $n^{(i)}$ columns of $\mathbf{X}$ that are associated with the subject (expression) label i.

First, for face (expression) recognition, the nuclear- (ℓ_1-) norm is chosen for $\mathbf{V}^{(i)}$ in the DICA, as the goal here is to uncover low-rank (sparse) components. Second, RPCA [9] is performed on each $\mathbf{X}^{(i)}$ for warm initialization of $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ in (5). Specifically, each basis $\mathbf{U}^{(i)}$ and component $\mathbf{V}^{(i)}$ is initialized as $\mathbf{U}^{(i)} = \mathbf{M}^{(i)}$ and $\mathbf{V}^{(i)} = \mathbf{M}^{(i)^T}$, respectively, where $\mathbf{A}^{(i)}$ denotes the low-rank matrix yielded by RPCA for subject (expression) i and $\mathbf{A}^{(i)} = \mathbf{M}^{(i)}\mathbf{\Sigma}\mathbf{N}^{(i)^T}$ denotes its skinny SVD. Note that setting $\mathbf{V}^{(i)} = \mathbf{M}^{(i)^T} = \mathbf{U}^{(i)^T}$ is an intuitive choice, considering that $\mathbf{V}^{(i)}$ and $\mathbf{U}^{(i)^T}$ are proportional to each other, as shown in Section II-A. Choosing an initial estimate that is close to the optimum sought can markedly speed up the convergence of a non-convex optimization problem like the DICA [47]. RPCA has been proved efficient in recovering low-complexity facial components, while also being robust to gross errors in the data [11]. This motivates its choice for the initialization step, while its positive impact on the convergence speed was corroborated by preliminary experiments. Third, Problem (5) is solved according to Algorithm 1.

Following a SRC-like approach, the class-specific reconstruction images $\{\mathbf{D}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}\}_{i=1}^{n_c}$ are concatenated to construct the dictionary $\mathbf{D}$. Then, for each query image $\mathbf{y} \in \mathbb{R}^{d\times 1}$ a vector $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{N\times 1}$ is sought so that $\mathbf{y}$ is represented as a sparse linear combination of the dictionary atoms, i.e., $\mathbf{y} = \mathbf{D}\hat{\boldsymbol{\alpha}}$. The sparse coefficient vector $\hat{\boldsymbol{\alpha}}$ is obtained by solving the Lasso minimization problem:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}\boldsymbol{\alpha}\|^2 + \lambda_{Lasso}\|\boldsymbol{\alpha}\|_1 \quad (12)$$

Finally, the subject (expression) label i^* is estimated as that accounting for the minimum class-specific reconstruction error of $\mathbf{y}$, i.e.,

$$i^* = \arg\min_{i\in\{1,2,\ldots,n_c\}} \left\|\mathbf{y} - \mathbf{D}\delta^{(i)}(\hat{\boldsymbol{\alpha}})\right\|, \quad (13)$$

where $\hat{\boldsymbol{\alpha}}$ is the solution of (12), and $\{\delta^{(i)}(\cdot) : \mathbb{R}^{N\times 1} \mapsto \mathbb{R}^{N\times 1}\}_{i=1}^{n_c}$ are class-specific selector operators calculated as

$$\delta^{(i)}(q_n) = \begin{cases} q_n, & \text{if } n \in \mathcal{S}^{(i)} \\ 0, & \text{otherwise} \end{cases} \quad (14)$$

Fig. 3. Example registered images from each of the 4 datasets used. From top to bottom: CK+ [51], AR [52], CMU Multi-PIE [53], GEMEP-FERA [54].

The proposed single-label framework is summarized in Algorithm 2 for face/expression recognition.

B. Multi-Label Case: Joint Face and Expression Recognition & Action Unit Detection

The framework described in the previous section is extended to the multi-label case, where each observation is associated with multiple labels w.r.t. different attributes. Two face analysis tasks that fall in this multi-label case are (a) joint face and expression recognition, and (b) facial action unit (AU) detection. In this section, we choose to present the DICA-based classification framework tailored to the aforementioned tasks, on which our experimental validation in Section IV is based.

Joint Face and Expression Recognition: First, the DICA (5) is solved for the total number of classes $n_c = n_s + n_e$, with n_s (n_e) being the number of subjects (expressions). Similarly to the single-label case, for the subject-(expression-)specific components $i \in \{1, 2, \ldots, n_s\}$ ($i \in \{n_s+1, n_s+2, \ldots, n_s+n_e\}$) the nuclear- ($\ell_1$-) norm is enforced on the corresponding $\mathbf{V}^{(i)}$. Second, the derived identity-related reconstruction images are used to form the identity dictionary $\mathbf{D}_\mathcal{I}$, while the expression-related reconstruction images are used to form the expression dictionary $\mathbf{D}_\mathcal{E}$. The final dictionary consists of the concatenation of $\mathbf{D}_\mathcal{I}$ and $\mathbf{D}_\mathcal{E}$ as $\mathbf{D} = \begin{bmatrix}\mathbf{D}_\mathcal{I} & \mathbf{D}_\mathcal{E}\end{bmatrix}$.

Subsequently, the SRC algorithm is modified accordingly to solve jointly for the identity and expression coefficient vectors $\hat{\boldsymbol{\alpha}}_\mathcal{I}$ and $\hat{\boldsymbol{\alpha}}_\mathcal{E}$, respectively:

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_\mathcal{I}, \hat{\boldsymbol{\alpha}}_\mathcal{E} &= \arg\min_{\boldsymbol{\alpha}_\mathcal{I}, \boldsymbol{\alpha}_\mathcal{E}} \frac{1}{2}\|\mathbf{y} - \begin{bmatrix}\mathbf{D}_\mathcal{I} & \mathbf{D}_\mathcal{E}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha}_\mathcal{I} \\ \boldsymbol{\alpha}_\mathcal{E}\end{bmatrix}]\|^2 \\ &+ \frac{\lambda_{Lasso}}{2}\| \begin{bmatrix}\boldsymbol{\alpha}_\mathcal{I} \\ \boldsymbol{\alpha}_\mathcal{E}\end{bmatrix} \|_1 \\ &= \arg\min_{\boldsymbol{\alpha}_\mathcal{I}, \boldsymbol{\alpha}_\mathcal{E}} \frac{1}{2}\|\mathbf{y} - \mathbf{D}_\mathcal{I}\boldsymbol{\alpha}_\mathcal{I} - \mathbf{D}_\mathcal{E}\boldsymbol{\alpha}_\mathcal{E}\|^2 \\ &+ \frac{\lambda_{Lasso}}{2}\|\boldsymbol{\alpha}_\mathcal{I}\|_1 + \frac{\lambda_{Lasso}}{2}\|\boldsymbol{\alpha}_\mathcal{E}\|_1\end{aligned} \tag{15}$$

Finally, the component separation approach of [11] is followed, where the reconstruction image $\hat{\mathbf{y}}_\mathcal{I} = \mathbf{D}_\mathcal{I}\hat{\boldsymbol{\alpha}}_\mathcal{I}$ based on the identity dictionary $\mathbf{D}_\mathcal{I}$ is utilized for face recognition, and, similarly, the reconstruction image $\hat{\mathbf{y}}_\mathcal{E} = \mathbf{D}_\mathcal{E}\hat{\boldsymbol{\alpha}}_\mathcal{E}$ based on the expression dictionary $\mathbf{D}_\mathcal{E}$ is utilized for expression recognition, according to the following minimum-residual rules:

$$i^*_\mathcal{I} = \arg\min_{i\in\{1,2,\ldots,n_s\}} \|\hat{\mathbf{y}}_\mathcal{I} - \mathbf{D}_\mathcal{I}\delta^{(i)}(\hat{\boldsymbol{\alpha}}_\mathcal{I})\| \tag{16}$$

$$i^*_\mathcal{E} = \arg\min_{i\in\{n_s+1,n_s+2,\ldots,n_s+n_e\}} \|\hat{\mathbf{y}}_\mathcal{E} - \mathbf{D}_\mathcal{E}\delta^{(i)}(\hat{\boldsymbol{\alpha}}_\mathcal{E})\| \tag{17}$$

In Fig. 2, one can see the decomposition of an expressive image into a identity-related component, an expression-related component and a sparse error term. The identity (expression) component is formed out of the reconstruction of the original image based on the corresponding subject- (expression-)specific subspace. It can be visually verified that indeed the identity (expression) component contains no expression-(subject-)related information, due to its calculation based on images of all training expressions (subjects) and the mutual incoherence property. Finally, the outliers term encodes whatever image features deviate in a non-Gaussian sense from the class-specific decomposition that model (5) dictates.

Facial Action Unit Detection: The DICA (5) is applied for the total of n_c of AU-specific classes, using the ℓ_1-norm to enforce sparse structure on the respective components $\{\mathbf{V}^{(i)}\}_{i=1}^{n_c}$. Note that a training image with more than one AUs activated can appear multiple times in (5), through the corresponding class-specific sub-matrices $\mathbf{X}_{\mathcal{S}^{(i)}}$. Similarly to Algorithm (2), reconstruction images are next used to form class-specific dictionaries $\mathbf{D}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}$, $i \in \{1, 2, \ldots, n_c\}$, each of which is associated only with the respective AU label, regardless of the possible presence of other AUs in the corresponding training images. The final dictionary $\mathbf{D} \in \mathbb{R}^{d\times N}$ is formed out of the concatenation of all class-specific dictionaries $\{\mathbf{D}^{(i)}\}_{i=1}^{n_c}$. Next, for each test set vector $\mathbf{y} \in \mathbb{R}^{d\times 1}$ the sparse coefficient vector $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{N\times 1}$ and the reconstructed test vector $\hat{\mathbf{y}} = \mathbf{D}\hat{\boldsymbol{\alpha}}$ are obtained by solving (12).

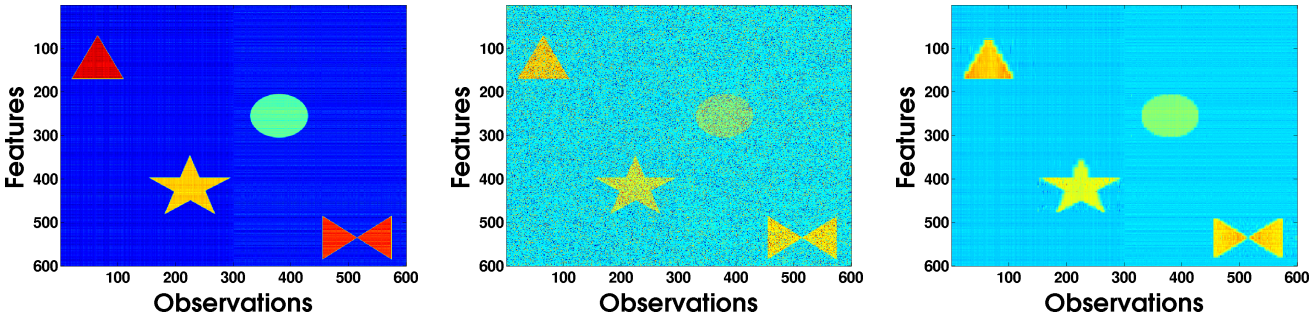
Classical SRC, formulated as in Equation (13), is not directly applicable to the action unit detection task, as the latter necessitates binary classification for each of the AU-specific classes. The sparse similarity voting approach in [55] is adopted herein for classification. Let $\boldsymbol{l_n} \in \{0,1\}^{n_c}$ be the binary label vector associated with the dictionary atom $\mathbf{d_n}$. By construction, only one element of $\boldsymbol{l_n}$ will be non-zero for our framework, i.e., that which corresponds to the AU label of the class-specific dictionary $\mathbf{d_n}$. Let also $\mathbf{L} \in \{0,1\}^{n_c\times N}$ be the label matrix for the whole dictionary, with corresponding label vectors $\boldsymbol{l_n}$ in its columns. Then, the multi-label *confidence* vector $\mathbf{c} \in \mathbb{R}^{n_c}$ for the test sample $\mathbf{y}$, is given by

$$\mathbf{c} = \sum_{n=1}^{N} w_n \boldsymbol{l_n} = \mathbf{L}\mathbf{w}\,, \tag{18}$$

where w_n denotes the similarity between the test vector $\mathbf{y}$ and its reconstruction by the n-th dictionary atom, given by

$$w_n = \frac{\hat{\alpha}_n \mathbf{d_n}^T\mathbf{y}}{\|\mathbf{y}\|\|\hat{\mathbf{y}}\|} \tag{19}$$

Each element c_i of the label vector $\mathbf{c}$ in (18) can be perceived as a *confidence* score with regards to the test sample belonging



(a) Synthetic Data (b) Data corrupted with sparse noise (c) Reconstruction by the DICA

Fig. 4. Illustration of corrupted synthetic data reconstruction, as produced by the DICA. Each 600×150 subset of the data matrix (where the first dimension is the feature space and the second dimension is the ambient space) is a superposition of one of the two low-rank components (depicted as 600×300 blue striped backgrounds in (a)) and one of the four block-sparse components, which form a shape of filled triangle, asterisk, circle and butterfly, respectively. (a) Original synthetic data, (b) Synthetic data of (a) contaminated with additive sparse noise, (c) Low-Rank/Sparse Reconstruction of the corrupted signal as produced by the DICA.

to the i-th AU class. Finally, binary labels for the test sample with respect to each class are obtained by thresholding each c_i via ROC analysis [56].

IV. Experiments

Our method is evaluated on four distinct tasks: (a) face recognition, (b) facial expression recognition, (c) joint face and expression recognition, and (d) facial action unit detection. Our dictionary-based framework for joint face and expression recognition is evaluated on CK+ Dataset [51], while experiments on subject-independent facial expression recognition are conducted on both CK+ [51] and CMU Multi-PIE [53] datasets. For face recognition experiments and action unit detection experiments, AR database [52] and GEMER-FERA database [54] is used, respectively.

The proposed method is compared to the approaches of Linear Regression Classifier (LRC) [57], Sparse Representation-based Classification (SRC) [15], as well as Robust Principal Component Analysis and SRC (RPCA+SRC) and Low-Rank Matrix Recovery with Structural Incoherence (LRSI) combined with SRC [12]. For RPCA+SRC, RPCA [9] is applied for each subject and the resulting low-rank (sparse) matrices are used for SRC-based face (expression) recognition similarly to [11]. For LRSI, the algorithm in [12] is applied subject-wise for face recognition and expression-wise for expression recognition; the nuclear norm is used for all components. In case of identical experimental protocol, LRSI results correspond to those reported in [12]. Unlike [12], where PCA is used to reduce dimensionality, vectorized images in the pixel domain are used for all experiments, with the exception of AU detection experiments in Section IV-E.

Implementation details: For both our method and LRSI, the parameter η that controls incoherence is set to the value 10^{-1}, which was proved efficient upon preliminary experiments. For the DICA, various values, different for each task, are examined for the parameter $\lambda^{(i)}$ controlling the norm $\|\mathbf{V}^{(i)}\|_{(\cdot)}$ and the outlier-related parameter λ_1 in Problem (5), and the best score achieved is reported each time. For each RPCA+SRC and LRSI optimization problem applied class-wise, the value $\lambda_1 = 1/\sqrt{\max(d, n^{(i)})}$ is used for the parameter associated with the sparse error term, which is an efficient heuristic according to [9].

TABLE I
Quantitative reconstruction results produced by the DICA on the synthetic data shown in Fig. 4. For a given component $\mathbf{X}^{(i)}$, the reconstruction metric used here corresponds to $\|\mathbf{X}^{(i)} - \hat{\mathbf{X}}^{(i)}\|_F / \|\mathbf{X}^{(i)}\|_F$, where $\hat{\mathbf{X}}^{(i)} = \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}$.

Reconstructions	
Clean Signal	0.369
Error Signal	0.916
Low-Rank Component 1	0.986
Low-Rank Component 2	0.972
Triangle	0.933
Asterisk	0.928
Circle	0.916
Butterfly	0.927
Relative Constraint	$9.9 \cdot 10^{-8}$

For the face recognition experiments in Section IV-C, the Lasso minimization problem (12) for the SRC-based approaches is solved by means of the Homotopy method [58], in order for our results to be comparable to those in [12]. For all SRC-based experiments in Sections IV-B, IV-D, and IV-E, the Efficient Euclidean Projections method [59] is chosen to solve the Lasso problems (12) and (15), thanks to its fast implementation and robustness to matrix singularities.

For all experiments with the DICA, the regularization parameter λ_{Lasso} of the Lasso minimization problems (12) and (15) is examined amongst the values $\{10^{-5}, 5 \cdot 10^{-5}, 10^{-4}, \ldots, 5 \cdot 10^{-1}\}$, and the best result is reported each time. For joint face and expression recognition, recognition accuracies reported correspond to the best average score over the two tasks. For all experiments with the other SRC-based approaches, that is, SRC, RPCA+SRC, and LRSI, λ_{Lasso} is fixed to 10^{-3}.

The DICA is also evaluated by means of experiments with synthetic data in Section IV-A. The results of these experiments serve as an important proof of concept since (a) they validate the effectiveness of our method both qualitatively and quantitatively, and (b) they provide evidence that our method

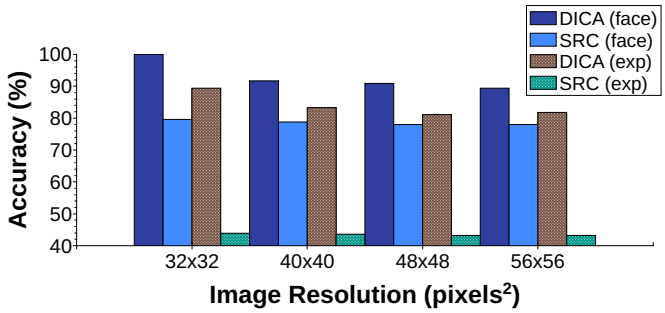


Fig. 5. Face and expression recognition accuracies (%), as produced by the DICA and SRC for the first fold of the protocol for the CK+ Dataset, varying with the image resolution.

can be applied equally well to any labeled data populations, thus serving diverse applications other than face analysis tasks.

A. Experiment on Synthetic Data

Our method is first evaluated on synthetic data corrupted with sparse, non-Gaussian noise. Each data point is constructed as a superposition of a low-rank and block-sparse component. In more detail, we first create a rank-2 component $\mathbf{X}^{(1)}$ with column space $\mathbf{U}^{(1)} \in \mathbb{R}^{600\times 2}$, based on the first two principal components of a random matrix $\mathbf{A} \in \mathbb{R}^{600\times 300}$. Next, we form a second rank-2 component $\mathbf{X}^{(2)}$ with column space $\mathbf{U}^{(2)} = \mathbf{R}\mathbf{U}^{(1)}$, where $\mathbf{R}$ is a random orthogonal matrix; as a result of this, the two components are mutually incoherent. Subsequently, four block-sparse components $\mathbf{X}^{(i)} \in \mathbb{R}^{600\times 150}$ $(3 \leq i \leq 6)$ are constructed, with their non-zero elements corresponding to visually discernible shapes, that is, triangle, asterisk, circle and butterfly, respectively. Those are then added to the low-rank components to form the matrices $\mathbf{Y}_1 = \mathbf{X}^{(1)} + \mathbf{X}^{(3)} + \mathbf{X}^{(4)}$ and $\mathbf{Y}_2 = \mathbf{X}^{(2)} + \mathbf{X}^{(5)} + \mathbf{X}^{(6)}$. Our final clean data matrix $\mathbf{Y}$ is the result of concatenation of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ along the second dimension, and can be seen in Fig. 4a.

Subsequently, sparse, non-Gaussian noise is added to the original signal $\mathbf{Y}$ to simulate a more realistic scenario. First, a matrix containing only values in $\{+1, -1\}$ is created as $\mathbf{E} = \text{sgn}(\mathbf{B})$, where $\mathbf{B} \in \mathbb{R}^{600\times 600}$ is a random matrix and sgn denotes the sign function. The final error matrix $\mathbf{O}$ is formed by setting to zero those entries of $\mathbf{E}$ whose indices i and j satisfy the rule $\mathcal{N}[i, j] \leq 0.8$, where $\mathcal{N} \in \mathbb{R}^{600\times 600}$ is a matrix whose elements follow the Normal distribution. The final corrupted signal $\tilde{\mathbf{Y}} = \mathbf{Y} + \mathbf{O}$ and the low-rank/sparse reconstruction produced by the DICA (5) can be seen in Fig. 4b and Fig. 4c, respectively. It is evident that our method reconstructs accurately all components, both the low-rank components lying in the background and the sparse components appearing as shapes, while, at the same time, isolates the sparse, gross errors. Quantitative results are reported in Table I, in terms of normalized reconstruction error for each component, that is, $\|\mathbf{X}^{(i)} - \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}^{(i)}\|_F / \|\mathbf{X}^{(i)}\|_F$. It is worth noting that all subspace-specific reconstruction errors along with the clean signal reconstruction error $\|\mathbf{Y} - \sum_{i=1}^{6} \mathbf{U}^{(i)}\mathbf{V}^{(i)}\mathbf{X}_{\mathcal{S}^{(i)}}\|_F / \|\mathbf{Y}\|_F$ have low value, corroborating the conclusions drawn for our method from the qualitative inspection of Fig. 4.

B. Joint Face & Expression Recognition on CK+ Dataset

Our method is evaluated on the two-label setting of joint face and expression recognition. CK+ [51] has been widely used for the task of face and posed expression recognition. It contains 123 subjects in a total of 593 sequences, 327 out of which are annotated with respect to the emotion portrayed. As our method does not consider the temporal dimension, only the last 4 frames are used as expressive images for each sequence, as those are close to the apex phase of the expression. The experimental setup is identical to that of [11]. Specifically, a subset of 25 subjects, corresponding to 108 sequences, is used herein that meet the following criteria: (a) there are at least 4 annotated sequences for each of them, and (b) they perform one of the 6 universal emotions[2](*Anger, Disgust, Fear, Happiness, Sadness and Surprise*). The first condition is essential in order for the subjects to appear with a sufficient amount of images in the training set (at least 12 images), and the resulting dictionary to be balanced (for the face recognition part). Example images for a female subject of CK+ can be seen in Fig. 3.

To examine how image dimensionality affects accuracy in both face and expression recognition and tune it accordingly, the following experiment is conducted. Specifically, the DICA and SRC are tested on joint face and expression recognition with the image resolution varying through the range 32×32, 40×40, 48×48 and 56×56 pixels. Note that all images have been previously converted to gray scale and aligned based on the location of the eyes. For each subject, 3 sequences are randomly picked to be used for training, leaving the rest for testing. The parameters of the DICA and SRC are optimized separately for each resolution and the best accuracy obtained is reported in Fig. 5. The choice of 32×32 pixels for the image size consistently leads to the best performance. This behaviour was expected as by using a smaller image size the *curse of dimensionality* is avoided (given that no feature extraction is performed to the aim of dimensionality reduction). It is also worth mentioning that using a smaller resolution for the DICA has the additional benefit of speeding-up the convergence, which increases quadratically with the dimensionality owing to the SVT operator (see Section II-B). Accuracies achieved using the three remaining resolutions do not vary largely. In view of the above, the image size is fixed to 32×32 pixels for all experiments of this section.

For joint face and expression recognition, for each subject, 3 sequences are randomly selected to be used for training, and the remaining sequences are used for testing. This process is repeated 10 times, and the average scores for the face and expression recognition tasks are reported. Leave-one-subject-out expression recognition experiments are also conducted and the average rate over 25 folds is reported. For all experiments, parameters $\lambda^{(i)}$ controlling the nuclear norm of the identity-related $\mathbf{V}^{(i)}$ in Problem (5) are set to 1. For joint face and expression recognition, the values for λ_1 and the expression-related $\lambda^{(i)}$ accounting for the best average score over the two tasks were found to be 10^{-2} and 10^{-2}, respectively. For

[2]18 sequences depicting 'Contempt' are not included.

TABLE II
RECOGNITION RATES (%) FOR JOINT FACE & EXPRESSION RECOGNITION AND SUBJECT-INDEPENDENT EXPRESSION RECOGNITION ON CK+ DATASET.

Method	Joint Face & Expression Recognition		Subject-Independent Expression Recognition
	Face	Expression	
LRC [57]	86.2	57.7	60.1
SRC [15]	75.4	41.4	53.5
RPCA+SRC [12]	89.6	59.5	70.6
LRSI [12]	92.9	75.5	71.4
DICA	96.7	83.6	75.7

expression recognition, the corresponding values were 10^{-2} and $5 \cdot 10^{-2}$, respectively.

Recognition rates for both tasks are reported in Table II. The merits of the DICA for face and expression recognition are directly evident from Table II: it is the best-performing method for both tasks, yielding face and expression recognition accuracies of 96.7% and 83.6%, respectively[3]. LRSI comes second in performance, by a negative margin of 3.8% and 8.1% for face and expression recognition, respectively. Surprisingly, LRC provides scores close to those obtained by RPCA+SRC, presumably due to the beneficial effect of small training size and the similarity between training and test data populations. It is worth stressing that results of the DICA and RPCA+SRC correspond to the same sparsity parameter $\lambda_{Lasso}/2$ being used for the two dictionaries in (15). We believe that by separately optimizing the sparsity parameters for the SRC coefficients of identity and expression classes, that is, $\boldsymbol{\alpha}_{\mathcal{I}}$ and $\boldsymbol{\alpha}_{\mathcal{E}}$, respectively, one can achieve even higher performance.

Our method achieves the best score of 75.7% in the second setup also, where facial expression is recognized on data from subjects unseen in the training phase. LRSI is again the second-best-performing method with 71.4%. SRC performs poorly in this setup too, primarily due to test images being associated with sparse linear combinations of similar faces rather than similar expressions in the dictionary.

Fig. 6 illustrates the low-rank identity-based reconstruction (Fig. 6b) and the sparse expression-based reconstruction (Fig. 6d), as produced by our method for the joint face and expression recognition experiment on CK+ images, grouped by subject (Fig. 6a) and by expression (Fig. 6c), respectively. Note that no expression variations are retained in the subject-based reconstruction, while, at the same time, the sparse expression components contain no subject-related information. It is also worth observing that the expression components (Fig. 6d) are 'denser' and also account for higher values in the image regions where the action units 'shaping' each corresponding expression lie [60] (e.g., Brow-Lowerer AU4 for 'Anger', or Lip Corner Depressor AU15 for 'Sadness'). Overall, the resulting reconstructions are discriminant for both tasks.

(a) (b) (c) (d)

Fig. 6. Joint Face and Expression Recognition on the CK+ Database: (a) Training images from six subjects showing various expressions, (b) Low-rank reconstruction produced by the DICA for each identity class, (c) Training images from six expression classes (from top to bottom: *Anger, Disgust, Fear, Happiness, Sadness, Surprise*) posed by various subjects, (d) Sparse reconstruction produced by the DICA for each expression class.

C. Face Recognition on AR Dataset

For the task of face recognition, the focus of experiments is to investigate methods' performances for varying percentage of face images corrupted due to occlusion in the training set. This is a frequently-occurring scenario in real-world biometrics applications, where noise-free training data is hard to be attained (e.g., due to uncontrolled recording conditions and huge amount of data). To this end, the AR Database [52] is used, which includes a total of 4,000 frontal images for 126 individuals. The face images exhibit variations with respect to expression, illumination and two types of occlusion, that is, sunglasses and scarf (see Fig. 3). For each subject, images are taken in two sessions, each one constituent of 13 images: 3 images with sunglasses, 3 with scarves, 4 with different expressions, and the remaining 3 with different illuminations. The latter 7 images, which do not include occlusions, are considered as neutral images for the experiments in this section.

A randomly picked subset of 100 subjects is used for our experiments. Three protocols are tested in an identical way as in [12], corresponding to occlusion in the training images due to (1) sunglasses, (2) scarf, and (3) sunglasses and scarf,

[3]The recognition scores obtained for the dictionary-based component separation (DCS) algorithm from [11] are 99.1% and 81.6% for joint face and expression recognition, respectively, and 86.8% for subject-independent expression recognition. These results are only to some extent comparable to those reported in Table II, given that the dataset and protocol are identical. However, bear in mind that in [11], K-SVD [38] is also applied to refine the identity and expression dictionaries, which are initially provided by RPCA [9]. For this reason, the corresponding results are not considered in the discussion of this section.

TABLE III
RECOGNITION RATES (%) FOR PROTOCOL 1 (SUNGLASSES) AND PROTOCOL 2 (SCARF) WITH VARYING PERCENTAGE OF OCCLUDED IMAGES ($n_o/7$) IN THE AR DATABASE TRAINING SET.

Method	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf	Sunglasses	Scarf
	$0\% = 0/7$		$14\% = 1/7$		$29\% = 2/7$		$43\% = 3/7$	
LRC [57]	61.3	59.5	69.2	66.7	72.9	73.3	73.3	73.2
SRC [15]	72.3	71.4	82.4	83.3	88.6	89.3	88.9	90.1
RPCA+SRC [12]	75.4	85.0	81.6	89.4	87.7	90.7	88.8	87.3
LRSI (reported in [12])	73.0	72.8	84.2	82.6	83.7	80.5	83.7	79.6
DICA	85.9	88.3	93.5	94.4	93.4	94.0	93.3	93.1

TABLE IV
RECOGNITION RATES (%) FOR PROTOCOL 3 (SUNGLASSES+SCARF) WITH VARYING PERCENTAGE OF OCCLUDED IMAGES ($2n_o/(7+2n_o)$) IN THE AR DATABASE TRAINING SET.

	Sunglasses+Scarf			
Method	$0\% = 0/(7+0)$	$22\% = 2/(7+2)$	$36\% = 4/(7+4)$	$46\% = 6/(7+6)$
LRC [57]	59.9	66.2	69.1	70.3
SRC [15]	71.6	82.1	89.0	90.5
RPCA+SRC [12]	72.5	86.3	90.8	93.1
LRSI (reported in [12])	62.8	80.8	81.8	82.8
DICA	81.8	93.8	95.2	95.4

respectively. Note that sunglasses account for occlusion of about 20% of the face image, whereas for the scarf scenario this percentage amounts to about 40%.

The three protocols are outlined below:

- **Protocol 1:** For each subject, n_{cl} neutral images and $n_o \in \{0, 1, 2, 3\}$ occluded images (sunglasses) from Session 1 are used for training, where $n_{cl}+n_o = 7$. 7 neutral images and 3 occluded images (sunglasses) from Session 2 are used for testing.
- **Protocol 2:** Same as Protocol 1, with occluded images containing scarf rather than sunglasses.
- **Protocol 3:** For each subject, $n_{cl} = 7$ neutral images, $n_{sg} \in \{0, 1, 2, 3\}$ sunglasses images, and $n_{sc} \in \{0, 1, 2, 3\}$ scarf images, from Session 1 are used for training, where $n_{sg} = n_{sc}$. Here, the amount of training images per subject varies from 7 to 13, as opposed to the first two protocols, in which it is fixed to 7. All 13 images (7 neutral, 3 sunglasses, 3 scarf) from Session 2 are used for testing.

Results are shown in Table III for Protocols 1 and 2, and in Table IV for Protocol 3. The DICA achieves the most accurate recognition in all scenarios, reaching 95.4% accuracy in Protocol 3 when 46% of training images are corrupted. The value of parameter λ_1 that yielded the best scores for our method was 10. It is worth noting that all methods show a significant increase in performance in all three protocols when at least one occluded image per subject is included in the training set, as compared to the case of 100% clean data. Notably, the performance achieved by the DICA fluctuates less as the percentage of training set corruption increases, as compared to that of the other methods. This is because components produced in the output of the DICA are by definition mutually incoherent, regardless of how many images with similar corruptions in similar face regions across classes are used for training. In Protocol 3, where two different kinds of data corruption are present, RPCA+SRC consistently achieves the second-best accuracy. It is also worth observing that, even for large percentages of training set corruption, SRC performs quite accurately also. This can be attributed to the efficiency of SRC in scenarios where the training and test set distributions are characterized by similar variations [61]. LRSI shows poor performance possibly due to its inability to suppress the effect of occlusion in the generated subspaces. LRC underperforms the rest of the methods in all cases. This can be largely attributed to singularities occurring in the matrix $\mathbf{D^T D}$, where $\mathbf{D}$ is the dictionary matrix (see [57], [61]).

In Fig. 7, the performance of our method and RPCA is comparatively illustrated on an instance of Protocol 1, that is, 7 images of a male subject, 3 of which are occluded by sunglasses. One can observe that both methods successfully remove variations caused by expression or illumination in the derived low-rank reconstruction. Nonetheless, our method succeeds to discard the occlusion in the reconstruction images, as opposed to the RPCA. This is due to the fact that presence of sunglasses in the reconstructed images of all subject classes would clash with the mutual incoherence property, which entails that class-specific components are as close as possible to being orthogonal. The same holds for Protocol 2, where the occlusion due to scarf covers even larger part of the image. Reconstructions yielded by our method for images of the same subject in Protocols 1 and 2 are shown in Fig. 8, for the scenario in which the occluded images cover $3/7$ of the training set.

D. Expression Recognition on CMU Multi-PIE Dataset

In Section IV-B we presented expression recognition experiments for the case of different subjects being included in the training and test set. Aiming to evaluate the effectiveness of our method in a scenario where labels from an additional source of variation, such as illumination, are not utilized in our discriminant analysis during training, we perform expression recognition also on the CMU Multi Pose Illumination, and

(a) Training Images (Protocol 1)

(b) Reconstruction by the RPCA

(c) Reconstruction by the DICA

Fig. 7. Face Recognition on the AR Database: Reconstruction images, as produced by the RPCA (b), and the DICA (c), on all training images of a subject in Protocol 1 (3/7=43% of occluded images (sunglasses)) (a).

(a) Training Images (Protocol 1)

(b) Reconstruction by the DICA

(c) Training Images (Protocol 2)

(d) Reconstruction by the DICA

Fig. 8. Face Recognition on the AR Database: Reconstruction produced by the DICA ((b),(d)), on all training images of a subject in Protocols 1 and 2 (3/7=43% of occluded images - sunglasses in (a) and scarf in (c), respectively).

Expression (Multi-PIE) Database [53]. This dataset contains 337 subjects, corresponding to about 750,000 images with 19 illumination variations, 15 different poses, and 6 facial expressions (*Neutral, Smile, Surprise, Disgust, Scream, Squint*). In the current study, only the frontal pose images are considered. For the presented experiments, 50 subjects are randomly selected. For each subject, 5 different illumination conditions are generated (corresponding to pan angles $-30°$, $-15°$, $0°$, $15°$, $30°$) for all 6 expressions, resulting in 30 images per subject. Some characteristic images from Multi-PIE are illustrated in Fig. 3.

The same protocol used in Section IV-B is adopted for facial expression recognition. Subject-independent experiments are conducted and the average score over 50 runs is reported. The best values for the sparsity-controlling parameters λ_1 and $\lambda^{(i)}$ for the expression components were found to be 10 and 1, respectively. Recognition rates are reported in Table V. Here, illumination conditions vary a lot across training images, rendering the task even more challenging. Still, our method achieves the best accuracy of 74.4%, followed by LRSI that achieves 67.3%. RPCA+SRC and SRC perform rather similarly, meaning that RPCA pre-processing fails in this case to uncover the class-specific low-rank manifolds. Note also that LRC shows a surprisingly poor performance. Again, the DICA efficiently decouples expression-related deformations from subject-specific characteristics and other effects, thereby enabling us to construct a much more discriminative expression dictionary.

TABLE V
RECOGNITION RATES (%) FOR SUBJECT-INDEPENDENT EXPRESSION RECOGNITION ON MULTI-PIE DATASET.

Method	Expression Recognition
LRC [57]	18.0
SRC [15]	58.9
RPCA+SRC [12]	60.4
LRSI [12]	67.3
DICA	74.4

E. Facial action unit detection on GEMEP-FERA Dataset

In this section, the efficiency of the DICA in decomposing an expressive image into mutually incoherent sparse components related to AUs is examined. The training subset of the GEMEP-FERA [54] dataset is used for subject-independent action unit detection experiments. It contains 7 subjects depicted in 87 image sequences, which are FACS-labeled on a frame-by-frame basis in terms of AUs. In this paper, we use only the images in which at least one out of 8 action units is activated. The AUs considered are: AU1 (Inner Brow Raiser), AU2 (Outer Brow Raiser), AU4 (Brow Lowerer), AU6 (Cheek Raiser), AU7 (Lid Tightener), AU12 (Lip Corner Puller), AU15 (Lip Corner Depressor), and AU17 (Chin Raiser). Images are converted to gray scale, aligned based on the location of the eyes, and, subsequently, resized to 128×128 pixels. Characteristic images are shown in Fig. 3. Intensities from 22×22 pixel patches around 15 facial points (extracted by the tracker in [62]) are gathered in a single vector for each image. The final feature vector is composed of PCA coefficients corresponding to components that account for 98% of the total variance (374 components in our experiments).

Seven-fold subject-independent cross-validation is performed, so that all images for the 7 subjects are tested. For each fold, a randomly selected subset of the training images, evenly distributed across subjects and AU labels, is used. For the DICA, the action unit detection framework described in Section III-B is used. Specifically, the rank $m^{(i)}$ of each subspace is set to 5, while the remaining parameters are optimized similarly to the previous experiments. The values of

TABLE VI
$F1$ SCORES (%) FOR EACH ACTION UNIT AND METHOD EXAMINED IN THE ACTION UNIT DETECTION EXPERIMENTS ON THE GEMEP-FERA DATABASE TRAINING SET.

AU	ML-kNN [63]	Rank-SVM [64]	LRC [57]	SRC [15]	STM [65]	DICA
1	53.8	67.0	37.7	60.5	68.1	66.3
2	41.6	46.3	47.2	58.3	65.5	58.9
4	20.3	20.4	61.5	58.3	43.3	55.5
6	62.2	68.2	57.1	63.9	71.6	70.2
7	53.7	61.9	66.2	67.8	66.2	70.6
12	75.8	77.7	75.8	76.9	82.1	78.0
15	28.1	44.8	46.2	30.1	39.3	41.0
17	39.3	38.0	18.6	37.8	35.9	32.0
Avg.	46.9	53.0	51.3	56.7	59.0	59.1

the sparsity-controlling parameters λ_1 and $\lambda^{(i)}$ accounting for the best performance were found to be 0.05 and 1, respectively.

Except for the DICA, LRC and SRC are also examined, while RPCA+SRC and LRSI are not considered, as their design is not adaptable to this task. Multi-Label k-Nearest Neighbours (ML-kNN) [63] ($k = 10$ neighbours) and Rank-SVM [64] (with polynomial kernel of degree 8) are also examined, as they are general-purpose algorithms for multi-label classification. For the DICA, each dictionary atom is associated with a single AU label (see Section III-B), as opposed to other methods, for which the training data retain their initial multi-class labelling. For the dictionary-based methods, namely the DICA, LRC and SRC, ROC ranking [56] is employed to threshold the class-specific *confidence* scores obtained by (18) and thus provide multi-class predictions for each test sample. Finally, for all algorithms examined in the experiments of this section, the $F1$ score, defined as $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$, is used as the evaluation metric.

Action unit detection results in terms of $F1$ score, as produced by each method, are reported in Table VI for each action unit along with the average performance over all AU classes. For comparison purposes, we choose to also include in Table VI the results reported in [65] for the same evaluation protocol for Selective Transfer Machine (STM), which is a recently published successful method for AU detection. The DICA achieves similar performance to that of STM[4], while it outperforms all other methods. SRC also achieves high performance, thus validating previous evidence that sparse representation is efficient for the AU detection task [27]. LRC, as well as the baseline methods ML-kNN and Rank-SVM, attain much poorer performance.

V. CONCLUSION AND FUTURE WORK

A method for recovering mutually incoherent and structured components in *face imagery*, relying on discriminant information as well as structure-inducing norms on the facial aspects, has been proposed in this paper. An ADMM-based algorithm that can solve appropriate minimization problems for the DICA, according to the matrix norm imposed, while also being robust to gross outliers through sparsity regularization, has been also proposed. Finally, a dictionary-based framework that combines the DICA with sparse representation to jointly address interrelated classification tasks within multi-label scenarios has been presented. The experimental validation of our method was primarily focused on face analysis tasks. The effectiveness of the DICA was first demonstrated on synthetic data contaminated with sparse, non-Gaussian noise. Next, extensive experiments were conducted on joint face and expression recognition, face recognition for varying percentages of corrupted images in the training set, subject-independent expression recognition under varying illumination conditions during training, as well as facial action unit detection. The DICA outperformed all methods that were used for comparison, in all tasks and experimental scenarios. Overall, the DICA is a robust framework that can generalize to classification of any number or type of labelled attributes that manifest themselves in the visual stream through specific structures, associated with mutually incoherent modes of variation.

Possible future research directions include the exploitation of alternative structures for component extraction induced by other matrix norms, the extension of the DICA to the temporal dimension, and its coupling with hierarchical/deep architectures, aiming at extracting incoherent, invariant subspaces.

[4]The difference in average performance over all AUs achieved by the DICA and the STM is not significant, according to a paired t-test at significance level 0.05.

ACKNOWLEDGMENTS

This work is funded by the EPSRC project EP/N007743/1 (FACER2VM). The work of C. Georgakis and Y. Panagakis is also partially supported by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 645094 (SEWA).

APPENDIX
SOLUTION OF PROBLEM (8)

Let us consider the problem (8). In this step of ADMM, we are minimizing w.r.t. $\mathbf{V}^{(i)}$ at iteration t, with $\{\mathbf{U}^{(i)}\}_{i=1}^{n_c}$, $\{\mathbf{V}^{(j)}[t]\}_{j\neq i}$, and $\mathbf{O}$ kept fixed. Let us re-write the problem for clarity of presentation:

$$\begin{aligned}\mathbf{V}^{(i)}[t+1] &= \arg\min_{\mathbf{V}^{(i)}} \mathcal{L}(\mathbf{V}^{(i)}, \mathbf{Y}[t], \mu[t]) \\ &= \arg\min_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \eta \sum_{i\neq j} \|\mathbf{V}^{(i)} \mathbf{V}^{(j)^T}\|_F^2 \\ &+ \frac{\mu[t]}{2} \|\mathbf{X} - \sum_{i=1}^{n_c} \mathbf{U}^{(i)} \mathbf{V}^{(i)} \mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O} + \mu[t]^{-1} \mathbf{Y}\|_F^2 \\ &= \arg\min_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + f(\mathbf{V}^{(i)})\end{aligned} \tag{20}$$

The minimizer (20) consists of a non-smooth term, induced by a norm function $\|\cdot\|_{(\cdot)}$, and a smooth, twice differentiable term described by the function f. It can easily be proved that the gradient ∇f is Lipschitz-continuous.

By linearizing f in the vicinity of the current point $\mathbf{V}^{(i)}[t]$, and by exploiting the Lipschitz-continuity of ∇f, we obtain the following equivalent problem

$$\begin{aligned}&\min_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + f(\mathbf{V}^{(i)}[t]) \\ &+ \mathrm{tr}\left(\nabla f(\mathbf{V}^{(i)}[t])^T (\mathbf{V}^{(i)} - \mathbf{V}^{(i)}[t])\right) \\ &+ \frac{L}{2} \|\mathbf{V}^{(i)} - \mathbf{V}^{(i)}[t]\|_F^2\end{aligned} \tag{21}$$

where $L > 0$ is an upper bound on the Lipschitz constant of ∇f. Problem (21) is re-written as

$$\min_{\mathbf{V}^{(i)}} \lambda^{(i)} \|\mathbf{V}^{(i)}\|_{(\cdot)} + \frac{1}{2}\|\mathbf{V}^{(i)} - (\mathbf{V}^{(i)}[t] - \frac{1}{L}\nabla f(\mathbf{V}^{(i)}[t])\|_F^2 \tag{22}$$

Having expressed the minimizer in this form, we now directly apply the SVT (shrinkage) operator, in case the nuclear- (ℓ_1-) norm is chosen for the first term of (22). For the nuclear norm, the solution is given by

$$\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{S}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - \frac{1}{L}\nabla f(\mathbf{V}^{(i)}[t])\right], \tag{23}$$

whereas for the ℓ_1-norm the solution is given by

$$\mathbf{V}^{(i)}[t+1] \leftarrow \mathcal{D}_{\lambda^{(i)}/L}\left[\mathbf{V}^{(i)}[t] - \frac{1}{L}\nabla f(\mathbf{V}^{(i)}[t])\right] \tag{24}$$

The gradient $\nabla f(\mathbf{V}^{(i)}[t])$ is computed as

$$\nabla f(\mathbf{V}^{(i)}[t]) = \left(-\mu[t]\mathbf{U}^{(i)}[t]^T\right)\left(\mathbf{X} - \sum_{i=1}^{n_c}\mathbf{U}^{(i)}[t]\mathbf{V}^{(i)}[t]\mathbf{X}_{\mathcal{S}^{(i)}} - \mathbf{O}[t] + \mu[t]^{-1}\mathbf{Y}[t]\right)\mathbf{X}_{\mathcal{S}^{(i)}}^T + 2\eta\sum_{j\neq i}\mathbf{V}^{(j)}[t]^T\mathbf{V}^{(j)}[t], \tag{25}$$

whereas an upper bound on the Lipschitz constant of ∇f is given by

$$L = 1.02\lambda_{\max}\left[\mu[t]\mathbf{X}_{\mathcal{S}^{(i)}}\mathbf{X}_{\mathcal{S}^{(i)}}^T + 2\eta\sum_{j\neq i}\mathbf{V}^{(j)}[t]^T\mathbf{V}^{(j)}[t]\right] \tag{26}$$

The respective closed-form solutions are obtained by substituting (25) and (26) into (23) or (24).

REFERENCES

[1] M. Pantic, "Facial Expression Analysis," in *The Encyclopedia of biometrics, S. Li and A. Jain, Eds.*, 2009, vol. 6, pp. 400–406.

[2] ——, *Facial Expression Recognition*, 2014, pp. 1–8.

[3] M. Turk, A. P. Pentland *et al.*, "Face recognition using eigenfaces," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1991, pp. 586–591.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[5] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using Laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005.

[6] L. K. Saul and S. T. Roweis, "Think globally, fit locally: unsupervised learning of low dimensional manifolds," *The Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[7] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[8] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[10] L. Ma, C. Wang, B. Xiao, and W. Zhou, "Sparse representation for face recognition based on discriminative low-rank dictionary learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2586–2593.

[11] S. Taheri, V. M. Patel, and R. Chellappa, "Component-Based Recognition of Faces and Facial Expressions," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 360–371, 2013.

[12] C.-P. Wei, C.-F. Chen, and Y.-C. F. Wang, "Robust Face Recognition With Structurally Incoherent Low-Rank Matrix Decomposition," *IEEE Transactions on Image Processing*, vol. 23, no. 8, 2014.

[13] Y. Panagakis, M. A. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust Correlated and Individual Component Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence, Special Issue in Multimodal Pose Estimation and Behaviour Analysis, (accepted)*.

[14] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "Subspace learning from image gradient orientations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 12, pp. 2454–2466, 2012.

[15] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[16] W. Deng, J. Hu, and J. Guo, "Extended SRC: Undersampled face recognition via intraclass variant dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1864–1870, 2012.

[17] E. Elhamifar and R. Vidal, "Robust classification using structured sparse representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1873–1879.

[18] W. Ou, X. You, D. Tao, P. Zhang, Y. Tang, and Z. Zhu, "Robust face recognition via occlusion dictionary learning," *Pattern Recognition*, vol. 47, no. 4, pp. 1559–1572, 2014.

[19] M. Yang, L. Zhang, S. C. Shiu, and D. Zhang, "Gabor feature based robust representation and classification for face recognition with Gabor occlusion dictionary," *Pattern Recognition*, vol. 46, no. 7, pp. 1865–1878, 2013.

[20] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Regularized robust coding for face recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1753–1766, 2013.

[21] J. Qian, J. Yang, F. Zhang, and Z. Lin, "Robust Low-Rank Regularized Regression for Face Recognition with Occlusion," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014, pp. 21–26.

[22] X. Jiang and J. Lai, "Sparse and Dense Hybrid Representation via Dictionary Decomposition for Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 5, pp. 1067–1079, 2015.

[23] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011)*. IEEE, 2011, pp. 827–834.

[24] S. Kaltwang, S. Todorovic, and M. Pantic, "Doubly Sparse Relevance Vector Machine for Continuous Facial Behavior Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2015, (to appear).

[25] C. Georgakis, S. Petridis, and M. Pantic, "Discrimination between native and non-native speech using visual features only," *IEEE Transactions on Cybernetics (T-CYB)*, 2015, (to appear).

[26] ——, "Discriminating native from non-native speech using fusion of visual cues," in *ACM Int'l Conf. on Multimedia (ACMMM)*, Orlando, Florida, USA, November 2014, pp. 1177–1180.

[27] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)* , 2011, pp. 336–342.

[28] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 2005, pp. II–370.

[29] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic, "A Dynamic Appearance Descriptor Approach to Facial Actions Temporal Modelling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, 2014.

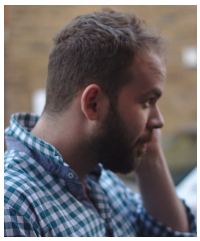
[30] P. Nagesh and B. Li, "A compressive sensing approach for expression-invariant face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1518–1525.

[31] S. Zafeiriou and M. Petrou, "Sparse representations for facial expressions recognition via l 1 optimization," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 32–39.

[32] R. Ptucha, G. Tsagkatakis, and A. Savakis, "Manifold based sparse representation for robust expression recognition without neutral subtraction," in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011, pp. 2136–2143.

[33] C.-S. Lee and R. Chellappa, "Sparse localized facial motion dictionary learning for facial expression recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 3548–3552.

[34] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.
[35] P. Ekman, W. Friesen, and J. Hager, "Facial action coding system." *Salt Lake City: Research Nexus eBook*, 2002.
[36] M. A. O. Vasilescu and D. Terzopoulos, "Multilinear subspace analysis of image ensembles," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, vol. 2, 2003, pp. II–93.
[37] H. Wang and N. Ahuja, "Facial expression decomposition," in *IEEE International Conference on Computer Vision*, 2003, pp. 958–965.
[38] M. Aharon, M. Elad, and A. Bruckstein, "The K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
[39] G. Papamakarios, Y. Panagakis, and S. Zafeiriou, "Generalised Scalable Robust Principal Component Analysis," in *British Machine Vision Conference (BMVC 2014)*, 9 2014.
[40] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Raps: Robust and efficient automatic construction of person-specific deformable models," in *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, June 2014.
[41] Y. Panagakis, C. L. Kotropoulos, and G. R. Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1905–1917, 2014.
[42] Y. Panagakis, M. Nicolaou, S. Zafeiriou, and M. Pantic, "Robust canonical time warping for the alignment of grossly corrupted sequences," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 540–547.
[43] P. Snape, Y. Panagakis, and S. Zafeiriou, "Automatic construction of robust spherical harmonic subspaces," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 218–233, 2015.
[44] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.
[45] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.
[46] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution," *Communications on pure and applied mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
[47] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.
[48] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.
[49] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2233–2246, 2012.
[50] G. Liu and S. Yan, "Active subspace: Toward scalable low-rank learning," *Neural computation*, vol. 24, no. 12, pp. 3371–3394, 2012.
[51] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 94–101.
[52] A. M. Martinez, "The AR face database," *CVC Technical Report*, vol. 24, 1998.
[53] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.
[54] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 42, no. 4, pp. 966–979, 2012.
[55] T. Sakai, H. Itoh, and A. Imiya, "Multi-label classification for image annotation via sparse similarity voting," in *Computer Vision–ACCV 2010 Workshops*. Springer, 2011, pp. 344–353.
[56] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
[57] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2106–2112, 2010.
[58] A. Y. Yang, S. S. Sastry, A. Ganesh, and Y. Ma, "Fast 1-minimization algorithms and an application in robust face recognition: A review," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 1849–1852.
[59] J. Liu and J. Ye, "Efficient Euclidean projections in linear time," in *ACM International Conference on Machine Learning*, 2009, pp. 657–664.
[60] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 1, pp. 28–43, 2012.
[61] D. Zhang, M. Yang, and X. Feng, "Sparse representation or collaborative representation: Which helps face recognition?" in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 471–478.
[62] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2013, pp. 896–903.
[63] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.
[64] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," in *Advances in neural information processing systems*, 2001, pp. 681–687.
[65] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 3515–3522.



Christos Georgakis received his Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens, Greece, in 2011. Currently, he is a member of the iBUG group, Department of Computing, Imperial College London, U.K., where he is pursuing the Ph.D. degree under the supervision of Prof. Maja Pantic. His research interests lie in human behaviour analysis, computer vision and statistical machine learning. He is a student member of the IEEE.

Yannis Panagakis is a Research Fellow in the Department of Computing, Imperial College London. He received his PhD and MSc degrees from the Department of Informatics, Aristotle University of Thessaloniki and his B.Sc. degree in Informatics and Telecommunication from the National and Kapodistrian University of Athens, Greece. Yannis received various scholarships and awards for his studies and research, including the prestigious Marie-Curie Fellowship in 2013. His current research interests include machine learning, signal processing, and mathematical optimization with applications to computer vision, human behaviour analysis, and audio analysis. He is a member of the IEEE.

Maja Pantic is a professor in affective and behavioral computing in the Department of Computing at Imperial College London, United Kingdom, and in the Department of Computer Science at the University of Twente, the Netherlands. She currently serves as the editor in chief of Image and Vision Computing Journal and as an associate editor for both the IEEE Transactions on Pattern Analysis and Machine Intelligence and the IEEE Transactions on Affective Computing. She has received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She is a fellow of the IEEE.